Acta Crystallographica Section A
**Foundations of Crystallography**
ISSN 0108-7673

Received 12 October 2001
Accepted 11 January 2002



# Direct methods and isomorphous replacement. The triplet invariant estimate when heavy atoms are located

**Carmelo Giacovazzo,[a,b]* Dritan Siliqi[c] and Liberato De Caro[a]**

[a]IRMEC–CNR, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania. Correspondence e-mail: c.giacovazzo@area.ba.cnr.it

The probabilistic theory of the three-phase structure invariants for isomorphous pairs has been generalized to the case in which a heavy-atom structure model is available. The rigorous method of joint probability distributions has been applied: it is able to handle errors in measurements and in the heavy-atom structure model, as well as the lack of isomorphism. The conclusive formulas have been successfully applied to experimental data.

## 1. Notation

$F_p = |F_p| \exp(i\phi_p)$: structure factor of the protein
$F_d = |F_d| \exp(i\phi_d)$: structure factor of the isomorphous derivative
$F_H = F_d - F_p$: structure factor of the heavy-atom structure (*i.e.* the atoms added to the native protein)
$\Phi_p = \phi_{p1} + \phi_{p2} + \phi_{p3}$: $p_1, p_2, p_3$ stand for $p\mathbf{h}_1, p\mathbf{h}_2, p\mathbf{h}_3$ with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$
$E_p = A_p + iB_p = R_p \exp(i\phi_p)$: normalized structure factor of the protein
$E_d = A_d + iB_d = R_d \exp(i\phi_d)$: pseudo-normalized structure factor of the derivative (normalized with respect to the native protein structure)
$E_H$: pseudo-normalized structure factor of the heavy-atom structure (normalized with respect to the native protein structure)
$\tau_i = \sum_{j=1}^{N} z_j^i$, $z_j$ = atomic number of the jth atom
$N_{\rm eq} = \tau_2^3/\tau_3^2$: (statistically equivalent) number of atoms in the primitive unit cell. $[N_{\rm eq}]_p$, $[N_{\rm eq}]_d$, $[N_{\rm eq}]_H$ refer to the protein, derivative and heavy-atom structure, respectively
$[\tau_2^3/\tau_3^2]_p$: value of $N_{\rm eq}$ for the native protein
$[\tau_2^3/\tau_3^2]_H$: value of $N_{\rm eq}$ for the heavy-atom structure
$[\tau_2^3/\tau_3^2]_d$: value of $N_{\rm eq}$ for the derivative
$f_j$: atomic scattering factor
$\sum_p = \sum_p f_j^2$: the sum is extended to the native protein atoms
$\sum_H = \sum_H f_j^2$: the sum is extended to the heavy atoms
$\sum_d = \sum_d f_j^2$: the sum is extended to the atoms in the derivative unit cell
$\Delta_{\rm iso} = |F_d| - |F_p|$

## 2. Introduction

The probability theory of the three-phase structure invariants for isomorphous pairs was initiated by Hauptman (1982; Hauptman *et al.*, 1982) who derived the joint probability distribution

$$P(\phi_{p1}, \phi_{p2}, \phi_{p3}, \phi_{d1}, \phi_{d2}, \phi_{d3}, R_{p1}, R_{p2}, R_{p3}, R_{d1}, R_{d2}, R_{d3}). \quad (1)$$

The Hauptman approach was revisited by Giacovazzo *et al.* (1988), who derived an efficient and simple formula for estimating the triplet phase $\Phi_p$. Their conclusive expression may be written as

$$P(\Phi_p | R_{pi}, R_{di}, i = 1, 2, 3) \approx [2\pi I_o(A)]^{-1} \exp(A \cos \Phi_p), \quad (2)$$

where

$$A = 2[N_{\rm eq}]_p^{-1/2} R_{p1} R_{p2} R_{p3} + 2[N_{\rm eq}]_H^{-1/2} \Delta_1 \Delta_2 \Delta_3 \quad (3a)$$
$$\Delta = R'_d - R'_p,$$
$$R'_p = |F_p| \Big/ \textstyle\sum_H^{1/2}, \quad R'_d = |F_d| \Big/ \textstyle\sum_H^{1/2}.$$

$R'_p = |E'_p|$ and $R'_d = |E'_d|$ are the structure-factor moduli of the protein and of the derivative, respectively, normalized with respect to the heavy-atom structure. The formulas (2) and (3*a*) were implemented into a direct-methods procedure aimed at phasing protein structure factors without any information on the heavy-atom positions (Giacovazzo, Cascarano, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo & Gonzales-Platas, 1995; Giacovazzo, Siliqi & Gonzales-Platas, 1995; Giacovazzo *et al.*, 1996).

When applied to real diffraction data, the procedure proved able to phase all the reflections up to derivative resolution and to provide, in favourable cases and in a completely automatic way, electron-density maps that may be directly interpreted. Severe lack of isomorphism between the native and the derivative may hinder the success: in this case, the electron-density maps are not straightforwardly interpretable but can still show a good correlation with the correct map.

One of the weak points of the Hauptman and Giacovazzo approaches is that they are unable to deal with errors in measurements and with the lack of isomorphism. A step in this direction has been made by Giacovazzo *et al.* (2001): their approach leads again to the von Mises distribution (2), but

$$A = 2[N_{\rm eq}]_p^{-1/2} R_{p1} R_{p2} R_{p3} + 2[N_{\rm eq}]_H^{-1/2} \prod_{i=1}^{3} \{\Delta_i/[1 + (\sigma_i^2)_H]\}, \tag{3b}$$

where $(\sigma_i^2)_H$ is the average square error normalized with respect to the heavy-atom substructure.

Fortier *et al.* (1985) and Klop *et al.* (1987) suggested a way for improving the electron-density maps obtainable *via* (1) and (2). Once the $\phi_p$ are approximately known, the heavy-atom structure is easily derivable *via* a difference Fourier synthesis with coefficients $(|F_d| - |F_p|)\exp(i\phi_p)$. The above authors proposed to incorporate the heavy-atom structure information into the triplet phase distribution *via* the doublet invariant $(\phi_{di} - \phi_{pi})$, $i = 1, 2, 3$. However, any application to real cases failed owing to the fact that distributions (1) and (2) were obtained in the absence of any information on $F_H$. Incorporating such information into (1) and (2) *a posteriori* [that is, after the mathematical form of (1) and (2) has been fixed on assuming $F_H$ unknown] is an unreliable way for improving their efficiency. Extensive tests made by Furey *et al.* (1990) suggest that the procedure is not able to eliminate the bias towards 'unresolved SIR values'. Further contributions by Fan & Gu (1985), Fan *et al.* (1990) and Liu *et al.* (1999) show that the bias may be overcome in favourable cases by a supplementary direct procedure combining Sim and Cochran distributions.

The most effective tool for checking the reliability of the triplet estimates when the heavy-atom structure is known should be the study of the joint probability density function

$$P(E_{p1}, E_{p2}, E_{p3}, E_{d1}, E_{d2}, E_{d3} | E_{H1}, E_{H2}, E_{H3}) \tag{4}$$

[from now denoted as $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ for shortness]. Unfortunately, $\mathbf{E}_p$, $\mathbf{E}_d$ and $\mathbf{E}_H$ are algebrically related by

$$E_{di} = E_{pi} + E_{Hi} \tag{5}$$

and therefore the distribution (4) would coincide with an unuseful Dirac-$\delta$-function-like shape [assuming zero values when (5) is not verified and unity when it is fulfilled].

The problem may be solved according to Giacovazzo & Siliqi (2001*a*,*b*): the experimental as well as the model errors are involved in the distributions like supplementary variables. Then (5) may be replaced by

$$E_{di} = E_{pi} + E_{Hi} + \sigma_i,$$

where $\sigma$ may represent any form of error; accordingly, distributions shaped as Dirac $\delta$ functions would no longer occur.

This paper is devoted to the study of the distribution (4) [that is $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$] under the assumption that errors of different nature affect the experimental $R_p$ and $R_d$ amplitudes as well as the $E_H$ values calculated from the heavy-atom structure model. We will also assume that:

(*a*) The atomic positions are the primitive random variables of our probabilistic approach.

(*b*)

$$F_{di} = F_{pi} + F_{Hi} + \mu_i \quad \text{for } i = 1, 2, 3, \tag{6}$$

where $\mu_j = |\mu_j| \exp(i\theta_j)$ is the cumulative error arising from errors in measurements, lack of isomorphism, errors in the heavy-atom structure *etc.* The notation throws the cumulative error on the derivative structure factor in accordance with Blow & Crick's (1959) treatment. In terms of $E$'s, (6) becomes

$$E_{di} = E_{pi} + E_{Hi} + \sigma_i \quad \text{for } i = 1, 2, 3, \tag{7}$$

so that

$$\langle |E_{di}|^2 \rangle = 1 + |E_{Hi}|^2 + |\sigma_i|^2 = |E_{Hi}|^2 + e_i,$$

where

$$e_i = 1 + |\sigma_i|^2 \quad |\sigma_i|^2 = |\mu_i|^2 \Big/ \sum_p.$$

(*c*) $\langle \mu_i \rangle = 0$ for $i = 1, 2, 3$.

(*d*) $\langle \mu_i \mu_j \rangle = 0$ for any pair $i$ and $j$ with $(i \neq j)$. This implies that errors are uncorrelated.

(*e*) Heavy-atom positions and native-protein-atom positions are uncorrelated, *i.e.* $\langle E_p E_H \rangle = 0$.

The above assumptions are often not completely fulfilled in practical cases (*i.e.* errors could be correlated) but are ideal for a first study of the problem.

## 3. The joint probability distribution $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ in $P\bar{1}$

Let us denote by

$$C(u_{p1}, u_{p2}, u_{p3}, u_{d1}, u_{d2}, u_{d3}) = \langle \exp[i(u_{p1}E_{p1} + u_{p2}E_{p2} + \ldots + u_{d3}E_{d3})] \rangle$$

(in short $C$) the characteristic function of the distribution $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ in $P\bar{1}$, where $u_{pi}, u_{di}, i = 1, 2, 3$, are carrying variables associated with $E_{pi}, E_{di}, i = 1, 2, 3$, respectively. We have (see Giacovazzo, 1998, ch. 5, for the technique)

$$C = \exp\left\{ i \sum_{j=1}^{3} u_{dj} E_{Hj} - \tfrac{1}{2}\left[ \sum_{j=1}^{3} (u_{pj}^2 + e_j u_{dj}^2 + 2u_{pj}u_{dj}) \right] - i[u_{p1}u_{p2}u_{p3} + u_{p1}u_{d2}u_{p3} + u_{p1}u_{p2}u_{d3} + u_{d1}u_{p2}u_{p3} + u_{d1}u_{d2}u_{p3} + u_{d1}u_{p2}u_{d3} + u_{p1}u_{d2}u_{d3} + u_{d1}u_{d2}u_{d3}]/[N_{\rm eq}]_p^{1/2} \right\}. \tag{8}$$

The joint probability distribution (4) is the Fourier transform of (8). We will adopt the following procedure: we will first expand $C$ in a Gram–Charlier series, we will then perform the Fourier transform of the series, and finally we will return back to the exponential form of the distribution. It has been shown by Giacovazzo & Siliqi (1996), for quartet as well for triplet invariants of two isomorphous structures, that such a procedure does not lose any information with respect to the practice of directly Fourier transforming (8). We have

$$\begin{aligned}P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H) &\approx (2\pi)^{-6} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \exp\left\{ -i\left[\sum_{j=1}^{3} u_{pj}E_{pj} + u_{dj}(E_{dj}^2 - E_{Hj})\right]\right. \\ &\left. - \tfrac{1}{2}\sum_{j=1}^{3}[u_{pj}^2 + e_j u_{dj}^2 + 2u_{pj}u_{dj}]\right\}\{1 - i[N_{\rm eq}]_p^{-1/2}[u_{p1}u_{p2}u_{p3} \\ &+ u_{p1}u_{d2}u_{p3} + u_{p1}u_{p2}u_{d3} + u_{d1}u_{p2}u_{p3} + u_{d1}u_{d2}u_{p3} \\ &+ u_{d1}u_{p2}u_{d3} + u_{p1}u_{d2}u_{d3} + u_{d1}u_{d2}u_{d3}]\}\,\mathrm{d}u_{p1}\ldots\mathrm{d}u_{d3}. \quad (9)\end{aligned}$$

Let us first integrate the component of order zero, and then the component of order $[N_{\rm eq}]_p^{-1/2}$. The integral of the first component may be written as (the bar on a matrix means 'transpose')

$$\begin{aligned}(2\pi)^{-6} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \exp(-i\overline{\mathbf{T}}\mathbf{U} - \tfrac{1}{2}\overline{\mathbf{U}}\mathbf{K}\mathbf{U})\,\mathrm{d}\mathbf{U} \\ = (2\pi)^{-3}[\det \mathbf{K}]^{-1/2}\exp(-\tfrac{1}{2}\overline{\mathbf{T}}\mathbf{K}^{-1}\mathbf{T}), \quad (10)\end{aligned}$$

where

$$\overline{\mathbf{U}} = [u_{p1}, u_{p2}, u_{p3}, u_{d1}, u_{d2}, u_{d3}],$$

$$\overline{\mathbf{T}} = [E_{p1}, E_{p2}, E_{p3}, E_{d1} - E_{H1}, E_{d2} - E_{H2}, E_{d3} - E_{H3}],$$

$$\mathbf{K} = \begin{vmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & e_1 & 0 & 0 \\ 0 & 1 & 0 & 0 & e_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & e_3 \end{vmatrix},$$

$$\det \mathbf{K} = \sigma_1^2\sigma_2^2\sigma_3^2.$$

The coefficients $\lambda_{ij}$ of $\mathbf{K}^{-1}$ are:

$$\begin{aligned}&\lambda_{11} = e_1/\sigma_1^2; \quad \lambda_{22} = e_2/\sigma_2^2; \quad \lambda_{33} = e_3/\sigma_3^3; \\ &\lambda_{44} = 1/\sigma_1^2; \quad \lambda_{55} = 1/\sigma_2^2; \quad \lambda_{66} = 1/\sigma_3^2; \\ &\lambda_{14} = \lambda_{41} = -1/\sigma_1^2; \quad \lambda_{25} = \lambda_{52} = -1/\sigma_2^2; \quad \lambda_{36} = \lambda_{63} = -1/\sigma_3^2.\end{aligned}$$

The other $\lambda_{ij}$ are identically zero.

In an explicit form, (10) may be rewritten as

$$(2\pi)^{-3}(\sigma_1\sigma_2\sigma_3)^{-1}\exp\left(-\tfrac{1}{2}\sum_{j=1}^{3}\{E_{pj}^2 + [E_{dj} - (E_{pj} + E_{Hj})]^2/\sigma_j^2\}\right). \quad (11)$$

The integral of the component of order $[N_{\rm eq}]_p^{-1/2}$ in (9) may be obtained by repeated application of the relation

$$\begin{aligned}\int_{-\infty}^{+\infty} (ix)^n \exp(-\beta^2x^2 - iqx)\,\mathrm{d}x \\ = 2^{-n}\pi^{1/2}\beta^{-n-1}\exp(-q^2/4\beta^2)H_n(q/2\beta)\end{aligned}$$

for $n = 0, 1$. $H_n$ is the Hermite polynomial of order $n$ [*i.e.* $H_0(x) = 1$, $H_1(x) = 2x$]. The final result is quoted in Appendix *A*. Then

$$\begin{aligned}P(\mathbf{E}_p, \mathbf{E}_d, |\mathbf{E}_H) &\approx (2\pi)^{-3}(\sigma_1\sigma_2\sigma_3)^{-1} \\ &\times \exp\left(-\tfrac{1}{2}\sum_{j=1}^{3}\{E_{pj}^2 + [E_{dj} - (E_{pj} + E_{Hj})]^2/\sigma_j^2\}\right. \\ &\left. + [N_{\rm eq}]_p^{-1/2}E_{p1}E_{p2}E_{p3}\right). \quad (12)\end{aligned}$$

The distribution (12) is the first main result of this paper. A qualitative analysis of the various terms in (12) suggests the following observations:

(*a*) each $|E_{pj}|$ is distributed according to the centric Wilson statistics [*i.e.* see the factor $\exp(-E_{pj}^2/2)$];

(*b*) the distribution of each $E_{dj}$ is centred about the value $(E_{pj} + E_{Hj})$: the larger $\sigma_j$, the sharper the distribution of $E_{dj}$ about $(E_{pj} + E_{Hj})$ will be (*i.e.* see the factor $\exp\{-\frac{1}{2}[E_{dj} - (E_{pj} + E_{Hj})]^2/\sigma_j^2\}$);

(*c*) the Cochran-type term (*i.e.* $\exp\{[N_{\rm eq}]_p^{-1/2}E_{p1}E_{p2}E_{p3}\}$), which agrees with the expected positivity of the triplet invariants of the protein, is the unique contributor of order $[N_{\rm eq}]_p^{-1/2}$. No terms of order $[N_{\rm eq}]_H^{-1/2}$ in (12) replace the contribution $[N_{\rm eq}]_H^{-1/2}\Delta_1\Delta_2\Delta_3$ in (3*a*), or $2[N_{\rm eq}]_H^{-1/2}\prod_{i=1}^{3}\{\Delta_i/[1 + (\sigma_i)_H^2]\}$ in (3*b*).

This result is very surprising but logically correct: the supplementary prior information on the $E_H$s generates very strong constraints of order zero on their distribution [see point (*b*)] which should not be modified by the influence of terms of order $[N_{\rm eq}]_H^{-1/2}$. Owing to the quite small value of $[N_{\rm eq}]_p^{-1/2}$, the Cochran contribution may be neglected in most cases. It should to be considered only when $[N_{\rm eq}]_p$ is small or when the $\sigma_i^2$ are quite large.

## 4. The triplet sign probability in $P\bar{1}$

Let $s_{p1}, s_{p2}, s_{p3}, s_{d1}, s_{d2}, s_{d3}$ be the signs of $E_{p1}, E_{p2}, \ldots, E_{d3}$, respectively. The calculation of the probability that $s_{p1}s_{p2}s_{p3} = 1$ requires various steps:

(*a*) the derivation of the marginal sign probability

$$P(s_{p1}, s_{p2}, s_{p3}|\mathbf{E}_H) = \sum_{s_{d1}, s_{d2}, s_{d3} = \pm 1} P(s_{p1}, s_{p2}, s_{p3}, s_{d1}, s_{d2}, s_{d3});$$

(*b*) the summation of the probability densities $P(s_{p1}, s_{p2}, s_{p3})$ over the combinations of the three signs $s_{p1}, s_{p2}, s_{p3}$ for which $s_{p1}s_{p2}s_{p3} = 1$ to obtain $P^+$;

(*c*) the summation of the probability densities $P(s_{p1}, s_{p2}, s_{p3})$ over the combinations for which $s_{p1}s_{p2}s_{p3} = -1$ to obtain $P^-$;

(*d*) the derivation of the normalized probability that the triplet sign is positive, *i.e.*

$$P_n^+ = (1 + P^-/P^+)^{-1}.$$

This procedure should lead to a rather complicated formula. Since

$$s_{di} \approx s_{pi}, \quad i = 1, 2, 3,$$

in most cases, we approximate $P(s_{p1}, s_{p2}, s_{p3})$ as follows:

$$P(s_{p1}, s_{p2}, s_{p3}) \approx L^{-1} \exp\left\{[N_{\rm eq}]_p^{-1/2} s_{p1}s_{p2}s_{p3}|E_{p1}E_{p2}E_{p3}| + \sum_{j=1}^{3} s_{pj}\Delta_{{\rm iso}j}F_{Hj}\Big/\mu_j^2\right\}, \quad (13)$$

where $L$ is a suitable normalization constant. We then perform the steps (*b*)–(*d*) of the procedure outlined above.

## 5. The joint probability distribution $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ in $P1$

As in the centric case, we will assume uncorrelated errors among the various $F_{dj}, j = 1, 2, 3$.

We will study first the characteristic function

$$C(u_{p1}, u_{p2}, u_{p3}, u_{d1}, u_{d2}, u_{d3}, {\rm v}_{p1}, {\rm v}_{p2}, \ldots, {\rm v}_{d3}), \quad (14)$$

where $u_{pi}, u_{dj}, {\rm v}_{pj}, {\rm v}_{dj}$ are carrying variables associated with $A_{pj}, A_{dj}, B_{pj}, B_{dj}$, respectively, for $j = 1, 2, 3$. Then the joint probability distribution [*i.e.* the Fourier transform of (14)] will assume the form

$$P(\mathbf{E}_p, \mathbf{E}_d, |\mathbf{E}_H) \approx \pi^{-6}(\sigma_1\sigma_2\sigma_3)^{-1}\prod_{i=1}^{3}(4R_{pi}R_{di}) \exp\left\{-\sum_{j=1}^{3}R_{pj}^2 - \sum_{j=1}^{3}|E_{dj} - (E_{pj} + E_{Hj})|^2/\sigma_j^2 + 2[N_{\rm eq}]_p^{-1/2}R_{p1}R_{p2}R_{p3}\cos(\phi_{p1} + \phi_{p2} + \phi_{p3})\right\}. \quad (15)$$

Considerations similar to those made in the centric case for (12) hold for (15) too. In particular: (*a*) $R_p$ is distributed according to the acentric Wilson statistics; (*b*) $E_{dj}$ is distributed around $E_{pj} + E_{hj}$: $\sigma_j$ defines the sharpness of the distribution; (*c*) the Cochran term is the only contribution of order $[N_{\rm eq}]_p^{-1/2}$; (*d*) no term of order $[N_{\rm eq}]_H^{-1/2}$ survives.

To derive the distribution of the protein phases given all the magnitudes, we will assume $\phi_{di} \approx \phi_{pi}$ for $i = 1, 2, 3$.

Then we will write

$$P(\phi_{p1}, \phi_{p2}, \phi_{p3}, |\mathbf{R}_p, \mathbf{R}_d, \mathbf{E}_H) \approx L^{-1}\exp\left\{\sum_{j=1}^{3}(2\Delta_{{\rm iso}j}|F_{Hj}|/|\mu_j|^2)\cos(\phi_{pj} - \phi_{Hj}) + 2[N_{\rm eq}]_p^{-1/2}R_{p1}R_{p2}R_{p3}\cos(\phi_{p1} + \phi_{p2} + \phi_{p3})\right\}, \quad (16)$$

where $L$ is a scale factor. Even if (16) has a simple form, the estimate of the protein triplet phase is not straightforward and some approximations are needed. Since $[N_{\rm eq}]_p^{-1/2}$ is usually very small, we introduce the following approach (see Giacovazzo, 1979):

(*a*) As a first approximation, the distribution (16) may be considered as the product of three statistically independent von Mises distributions

$$M(\phi_{pj}; \phi_{Hj}, G_j) \approx L^{-1}\exp[G_j\cos(\phi_{pj} - \phi_{Hj})], \quad j = 1, 2, 3 \quad (17)$$

with $G_j = 2\Delta_{\rm iso}|F_{Hj}|/|\mu_j|^2$.

(*b*) Each $M(\phi_{pj}; \phi_{Hj}, G_j)$ may be approximated (Stephens, 1963) by the wrapped normal distribution $W_N[D_1(G_j), \phi_{Hj}]$, where

$$W_N(\rho, \theta) = \left[1 + 2\sum_{p=1}^{\infty}\rho^{p^2}\cos p(\phi - \theta)\right]\Big/2\pi$$

with $0 < \phi \leq 2\pi$, $0 \leq \rho \leq 1$, $\rho = \exp(-\sigma^2/2)$ and $D_1(x) = I_1(x)/I_0(x)$ is the ratio of the modified Bessel functions of order one and zero, respectively.

(*c*) The convolution of the three von Mises distributions (17) (providing the estimate of $\Phi_p$) may be replaced by the convolution of the three wrapped normal distributions which is equal to

$$W_N[D_1(G_1)D_1(G_2)D_1(G_3), \phi_{H1} + \phi_{H2} + \phi_{H3}]. \quad (18)$$

(*d*) Equation (18) may in turn be approximated by the von Mises distribution

$$M(\Phi_p; \Phi_H, T) \approx L^{-1}\exp[T\cos(\Phi_p - \Phi_H)], \quad (19)$$

where

$$\Phi_H = \phi_{H1} + \phi_{H2} + \phi_{H3}$$

and $T$ is defined *via* the relationship

$$D_1(T) = D_1(G_1)D_1(G_2)D_1(G_3). \quad (20)$$

(*e*) Even if (19) is a good approximation of $P(\Phi_p|\mathbf{R}_p, \mathbf{R}_d, \mathbf{E}_H)$, we can combine it with the Cochran contribution in (16). By considering the Cochran term as statistically independent of (19) (indeed the first depends on the moduli $R_{pj}$, the second on the differences $\Delta_{{\rm iso}j}$), we have

$$P(\Phi_p|\mathbf{R}_p, \mathbf{R}_d, \mathbf{E}_H) \approx L^{-1}\exp[G\cos(\Phi_p - \Theta_p)], \quad (21)$$

where

$$G = (T^2 + C^2 + 2CT\cos\Phi_H)^{1/2}$$
$$C = 2[N_{\rm eq}]_p^{-1/2}R_{p1}R_{p2}R_{p3}$$
$$\tan\Theta_p = T\sin\Phi_H/(T\cos\Phi_H + C).$$

$\Theta_p$ is the most probable phase of $\Phi_p$ and may vary in the interval $(0, 2\pi)$, $G$ is its reliability parameter.

## 6. Experimental applications

The unexpected result contained in (12) and (15) (*i.e.* no term of order $N_H^{-1/2}$ survives) suggests that the final probability distribution (21) should be strictly correlated with classical SIR techniques. It is then mandatory to check this result *via* experimental applications. We used two test structures:

**Table 1**
BPO, calculated data.

Average phase errors for the 23955 measured reflections obtained at the end of the tangent process based on equation (21); two, four and six Au sites were assumed. Corresponding errors obtained *via* classical SIR techniques are also quoted.

| Au sites | Equation (21) $\langle\lvert\Delta\phi\rvert^\circ\rangle$ | SIR $\langle\lvert\Delta\phi\rvert^\circ\rangle$ |
|---|---|---|
| 2 | 44 | 44 |
| 4 | 45 | 46 |
| 6 | 47 | 47 |

**Table 2**
BPO, calculated data.

These data were computed by locating two Au atoms in the natural sites. To simulate errors in the heavy-atom model we have recalculated the $F_H$'s after average atomic shifts $\langle d\rangle = 0.05, 0.5, 0.5, 1$ Å.

| $\langle d\rangle$ | Equation (21) $\langle\lvert\Delta\phi\rvert^\circ\rangle$ | SIR $\langle\lvert\Delta\phi\rvert^\circ\rangle$ |
|---|---|---|
| 0.05 | 45 | 45 |
| 0.2 | 48 | 48 |
| 0.5 | 57 | 57 |
| 1 | 74 | 74 |

(i) M-FABP (Zanotti *et al.*, 1992), space group $P2_12_12_1$, $a$ = 35.9, $b$ = 56.5, $c$ = 72.7 Å, molecular formula $C_{667}N_{170}O_{261}S_3$, one formula per asymmetric unit, data resolution of 2.14 Å for the native and for the Hg derivative (7595 measured reflections), one Hg site in the asymmetric unit.

(ii) BPO (Hecht *et al.*, 1994), space group $P2_13$, $a$ = 126.5 Å, molecular formula $C_{2744}O_{1073}N_{712}$, one formula per asymmetric unit, data resolution 2.35 Å for the native (23956 measured reflections), 2.8 Å resolution for the Au derivative (15741 measured reflections), two Au sites in the asymmetric unit.

For M-FABP and BPO, the program *MLPHARE* (Collaborative Computational Project, Number 4, 1994) calculated, for the acentric reflections, phasing power values equal to 1.14 and 1.10, respectively.

We first used calculated data. We added some Au sites to the 'natural sites' of the BPO derivative to check how the accuracy of (21) changes with the scattering power of the heavy-atom structure. We show in Table 1 the average phase errors for the 23955 reflections [as obtained after the application of the tangent formula based on (21)] when two, four and six Au sites are located; the corresponding phase errors obtained *via* classical SIR techniques are also quoted.

To find the accuracy limits of (21), we used again the calculated data of BPO, with two Au sites in the 'natural' positions. To simulate errors into the heavy-atom structure model, we used in (21) the calculated $F_H$ values after having moved the two Au atoms by average shifts of 0.05, 0.2, 0.5 and 1 Å, respectively. The average phase errors for the 23955 reflections are shown in Table 2 and are compared with the corresponding values obtained *via* classical SIR techniques.

**Table 3**
Cumulative average phase error $\langle\lvert\Delta\Phi\rvert^\circ\rangle$ for equations (2) and (21).

NTR is the number of triplets with $|A|$ or $G$ larger than the threshold value TRS.

M-FABP

| TRS | $NTR_{(2)}$ | $\langle\lvert\Delta\Phi\rvert^\circ\rangle_{(2)}$ | $NTR_{(21)}$ | $\langle\lvert\Delta\Phi\rvert^\circ\rangle_{(21)}$ |
|---|---|---|---|---|
| 0.4 | 50.000 | 72 | 50.000 | 68 |
| 1.2 | 26.522 | 69 | 35.540 | 67 |
| 2.0 | 4.076 | 64 | 18.715 | 65 |
| 3.8 | 76 | 48 | 3.534 | 56 |
| 5.5 | 0 | – | 754 | 50 |

BPO

| TRS | $NTR_{(2)}$ | $\langle\lvert\Delta\Phi\rvert^\circ\rangle_{(2)}$ | $NTR_{(21)}$ | $\langle\lvert\Delta\Phi\rvert^\circ\rangle_{(21)}$ |
|---|---|---|---|---|
| 0.4 | 50.000 | 72 | 50.000 | 55 |
| 0.8 | 41.862 | 71 | 17.100 | 55 |
| 1.2 | 9.804 | 64 | 4.493 | 54 |
| 1.6 | 2.327 | 56 | 847 | 51 |
| 2.0 | 41 | 43 | 7 | 38 |

**Table 4**
For each test structure and for each procedure, NREF is the number of phased reflections, $\langle\lvert\Delta\phi\rvert^\circ\rangle$ and $\langle(\lvert\Delta\phi\rvert^\circ)_w\rangle$ are the average and the weighted average error, respectively, CORR the correlation factor of the corresponding electron-density map.

M-FABP

| | NREF | $\langle\lvert\Delta\phi\rvert^\circ\rangle$ | $\langle(\lvert\Delta\phi\rvert^\circ)_w\rangle$ | CORR |
|---|---|---|---|---|
| MLPHARE | 7.040 | 69 | (57) | 0.41 |
| Equation (2) | 7.121 | 70 | (62) | 0.41 |
| Equation (21) | 7.031 | 67 | (61) | 0.41 |

BPO

| | NREF | $\langle\lvert\Delta\phi\rvert^\circ\rangle$ | $\langle(\lvert\Delta\phi\rvert^\circ)_w\rangle$ | CORR |
|---|---|---|---|---|
| MLPHARE | 15.741 | 61 | (49) | 0.47 |
| Equation (2) | 15.722 | 63 | (53) | 0.44 |
| Equation (21) | 15.046 | 63 | (51) | 0.44 |

The last step of our calculations is addressed to check the relative efficiency of (2) and (21) by experimental data. We first estimated the triplet phase invariants found among the NLARGE reflections with the largest value of $|\Delta_{iso}|$ (NLARGE = 920 for M-FABP, NLARGE = 840 for BPO). In Table 3, $\langle\lvert\Delta\Phi\rvert^\circ\rangle$ is the cumulative average phase error of the triplets with $G \geq$ TRS if (21) is used, with $|A| \geq$ TRS if (2) is used. Eqations (2) and (21) are tools for phase assignment and extension through appropriate tangent formulas. To check if (21) leads to better phase estimates (for single reflections) than (2), we used both in the tangent procedure described by Giacovazzo *et al.* (1996), aiming at phasing reflections up to derivative resolution. The quality of the phases has been monitored by calculating the correlation factor (CORR) between the corresponding electron density $\rho$ and the 'correct' map $\rho_{mod}$ (obtained *via* model phases, all the reflections up to native resolution included):

$$\mathrm{CORR} = (\langle\rho\rho_{\mathrm{mod}}\rangle - \langle\rho\rangle\langle\rho_{\mathrm{mod}}\rangle) \times [(\langle\rho^2\rangle - \langle\rho\rangle^2)^{1/2}(\langle\rho^2_{\mathrm{mod}}\rangle - \langle\rho_{\mathrm{mod}}\rangle^2)^{1/2}]^{-1}.$$

To have a reference standard, the phases were also determined *via* the classical SIR techniques by using the program *MLPHARE* (Collaborative Computational Project, Number 4, 1994)

The results are shown in Table 4. We note:

(*a*) Tables 1 and 2 show that the procedures phasing protein reflections *via* triplets and *via* classical SIR techniques have equivalent accuracy;

(*b*) the larger efficiency shown in Table 3 by (21) with respect to (2) in phasing the triplet invariants found among the NLARGE reflections is not confirmed when the formula is applied to all the measured reflections. Indeed, the final weighted average phase error in Table 4 is almost equal for the two equations and for both the test structures.

(*c*) MLPHARE provides results equivalent to those obtained *via* (2) and (21) for M-FABP, while it is slightly more efficient for BPO. This is probably due to the maximum-likelihood refinement techniques used to refine heavy-atom parameters [a rather naive least-squares techniques is used in our procedure when (21) is used].

## 7. Conclusions

The study of the distributions $P(\mathbf{E}_p, \mathbf{E}_d)$ and $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ leads in both the cases to von Mises distributions. The concentration parameter of $P(\mathbf{E}_p, \mathbf{E}_d)$ contains a term of order $N_H^{-1/2}$ [see (2)], which establishes the correlation of the triplet method with standard SIR techniques (Giacovazzo *et al.*, 1996). The concentration parameter of $P(\mathbf{E}_p, \mathbf{E}_d|\mathbf{E}_H)$ directly and uniquely relies on the probability parameters of the SIR technique. It may be concluded that the three methods are practically equivalent, the triplet method having the advantage of phasing reflections in complete automation, without previous knowledge of the heavy-atom structure. The practical suggestion coming from this paper is the following three-step phasing procedure: (i) a first batch of reflections (say the NLARGE subset) is phased *via* (2); (ii) heavy-atom positions are found *via* a differential Fourier synthesis and refined *via* suitable least squares; (iii) the phasing process is extended to smaller $|E|$s *via* SIR–MIR techniques in order to save computing time and storage (the millions of triplet invariants necessary to the entire phasing process are no longer calculated). This mixed procedure still offers the complete automation of the process (from experimental data to phased protein reflections) and may be automatically connected with phase-refinement procedures like solvent flattening and/or histogram matching.

## APPENDIX *A*

The integral of the component of order $[N_{\mathrm{eq}}]_p^{-1/2}$ in (9) may be written as follows:

$$\begin{aligned}
&(2\pi)^{-3}(\sigma_1\sigma_2\sigma_3)^{-1}\exp\left(-\tfrac{1}{2}\sum_{j=1}^{3}\{E_{pj}^2 + [E_{dj} - (E_{pj} + E_{Hj})]^2/\sigma_j^2\}\right)\\
&\times \sigma_1^2\sigma_2^2\sigma_3^2[N_{\mathrm{eq}}]_p^{-1/2}\{[E_{p1}e_1 - (E_{d1} - E_{H1})]\\
&\times [E_{p2}e_2 - (E_{d2} - E_{H2})][E_{p3}e_3 - (E_{d3} - E_{H3})]\\
&+ [E_{p1}e_1 - (E_{d1} - E_{H1})][E_{d2} - (E_{p2} + E_{H2})]\\
&\times [E_{p3}e_3 - (E_{d3} - E_{H3})] + [E_{p1}e_1 - (E_{d1} - E_{H1})]\\
&\times [E_{p2}e_2 - (E_{d2} - E_{H2})][E_{d3} - (E_{p3} + E_{H3})]\\
&+ [E_{d1} - (E_{p1} + E_{H1})][E_{p2}e_2 - (E_{d2} - E_{H2})]\\
&\times [E_{p3}e_3 - (E_{d3} - E_{H3})] + [E_{d1} - (E_{p1} + E_{H1})]\\
&\times [E_{d2} - (E_{p2} + E_{H2})][E_{p3}e_3 - (E_{d3} - E_{H3})]\\
&+ [E_{d1} - (E_{p1} + E_{H1})][E_{p2}e_2 - (E_{d2} - E_{H2})]\\
&\times [E_{d3} - (E_{p3} + E_{H3})] + [E_{p1}e_1 - (E_{d1} - E_{H1})]\\
&\times [E_{d2} - (E_{p2} + E_{H2})][E_{d3} - (E_{p3} + E_{H3})]\\
&+ [E_{d1} - (E_{p1} + E_{H1})][E_{d2} - (E_{p2} + E_{H2})]\\
&\times [E_{d3} - (E_{p3} + E_{H3})]\}
\end{aligned} \tag{22}$$

A long but trivial algebraic analysis shows that the linear part of (22) reduces to $[N_{\mathrm{eq}}]_p^{-1/2}E_{p1}E_{p2}E_{p3}$.

We thank the referees for useful suggestions and discussions.

## References


Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Fan, H.-F. & Gu, Y.-X. (1985). *Acta Cryst.* A**41**, 280–284.
Fan, H.-F., Hao, Q., Gu, Y.-X., Qian, J.-Z. & Zheng, C.-D. (1990). *Acta Cryst.* A**46**, 935–939.
Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.
Furey, W. Jr, Chandrasekhar, K., Dyda, F. & Sax, M. (1990). *Acta Cryst.* A**46**, 560–567.
Giacovazzo, C. (1979). *Acta Cryst.* A**35**, 757–764.
Giacovazzo, C. (1998). *Direct Phasing in Crystallography*. Oxford University Press.
Giacovazzo, C., Cascarano, G., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503–510.
Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst.* A**44**, 45–51.
Giacovazzo, C. & Gonzales-Platas, J. (1995). *Acta Cryst.* A**51**, 398–404.
Giacovazzo, C. & Siliqi, D. (1996). *Acta Cryst.* A**52**, 133–142.
Giacovazzo, C. & Siliqi, D. (2001*a*). *Acta Cryst.* A**57**, 40–46.
Giacovazzo, C. & Siliqi, D. (2001*b*). *Acta Cryst.* A**57**, 414–419.
Giacovazzo, C., Siliqi, D. & Garcia-Rodriguez, L. (2001) *Acta Cryst.* A**57**, 571–575.
Giacovazzo, C., Siliqi, D. & Gonzales-Platas, J. (1995). *Acta Cryst.* A**51**, 811–820.
Giacovazzo, C., Siliqi, D., González-Platas, J., Hecht, H., Zanotti, G. & York, B. (1996). *Acta Cryst.* D**52**, 813–825.
Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609–621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* A**51**, 177–188.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.
Hecht, H., Sobek, H., Haag, T., Pfeifer, O. & Van Pee, K. H. (1994). *Nature Struct. Biol.* **1**, 532–537.
Klop, E. A., Krabbendam, H. & Kroon, J. (1987). *Acta Cryst.* A**43**, 810–820.
Liu, Y.-D., Harvey, I., Gu, Y.-X., Zheng, C.-D., He, Y.-Z., Fan, H.-F., Hasnain, S. S. & Hao, Q. (1999). *Acta Cryst.* D**55**, 1620–1622.
Stephens, H. A. (1963). *Biometrika*, **50**, 385–390.
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.